Clustering Districts in Paris

- including relationship with price of apartment

Seohyun PARK

1. Introduction

Even if we call Paris, all of its districts, different districts are filled with numerous kinds of venues which define their different environment. It is possible to segment the different venues in a neighborhood according to category, and then to group the neighborhoods together which have similar kinds of neighborhoods. This method may serve as a variable to help make a decision when people consider moving to Paris. Also, with the comparison of the price of apartment in each district, we can find out an alternative district which is less expensive than a certain district.

2. Data Acquisition and Cleaning

This project works with a data set. The dataset consists of geometric coordinates of each district of Paris which I found on a website and I got it by Scrapping. By using BeautifulSoup, I could convert it to csv data type, and finally resulted in a dataframe which consists of 4 columns : PostalCode, Price/m2, Latitude and Longitude. With geometric coordinates for Paris, I could get different venues by Foursquare API. It provides access to an enormous database consisting of venues from all around the world.

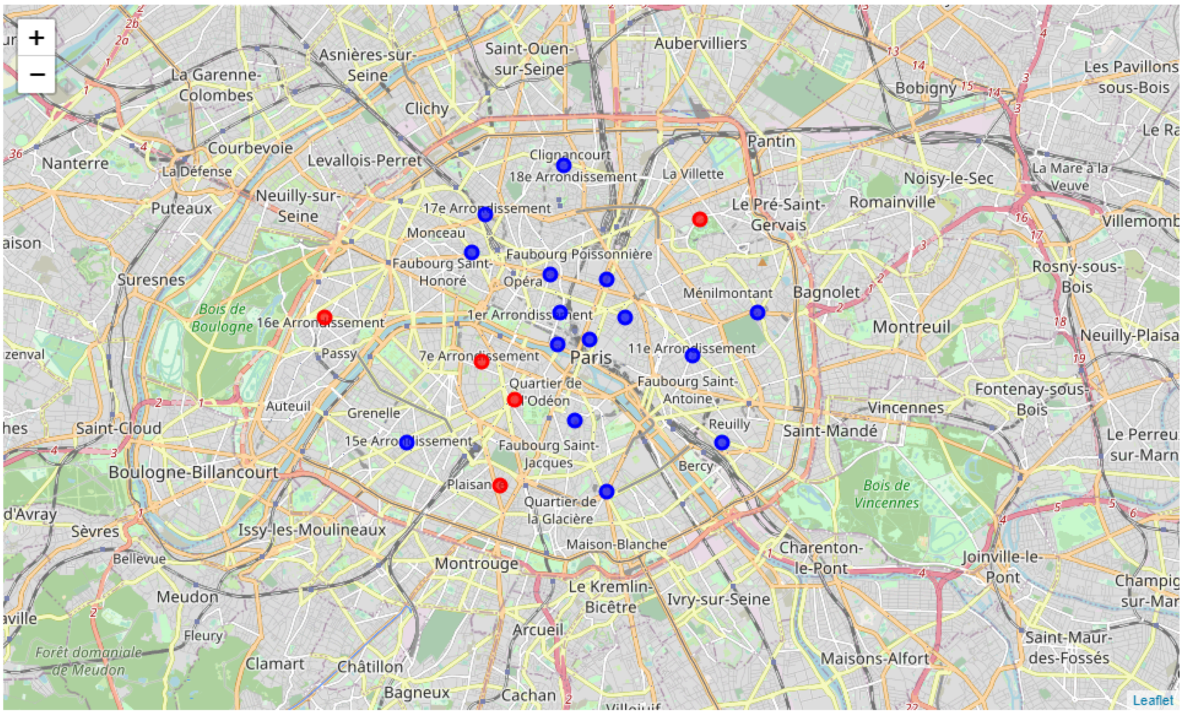
3. Methodology

The goal of this project is first of all to group together the similar neighborhoods in districts of Paris. After the dataset is ready by Foursquare API which demonstrates different venues in each district, I could find out which type of venue is the most present in each district. For grouping by these features, I decided to use K-means. K-means is one of the machine learning algorithms that groups a dataset into a user-specified number of clusters. When we use K-means clustering, we need some way to determine the number of clusters. How can we determine it ? The elbow method is to run k-means clustering on the dataset for a range of values of k, and for each value of k it calculates the sum of squared errors. And then, we choose the smallest value of k which represents where we start to have diminishing returns by increasing k. There is another method except elbow method, it is silhouette analysis. It is a way to measure how close each point in a cluster is to the points in its neighboring clusters and to find out the optimum value for k during k-means clustering. The value is between -1 and 1, if the value is close to 1, it means that the sample is far away from its neighboring cluster and very close to the cluster it is assigned. As the same, if the value is close to -1, it means that the sample is closer to its neighboring cluster than to the cluster it is assigned. How about if the sample has 0 value ? it means that it is in the middle of the two clusters.

4. Result

4-1. Similarity of neighborhood

In this project, we decide to set the value of cluster as 2 for dividing districts by 2 : Residential area and commercial area.



Districts pinned on Paris after clustering

The red point represents the residential area : 6th, 7th, 14th, 16th and 19th districts. And the blue point represents the commercial area : 1er, 2nd, 3rd, 4th, 5th, 8th, 9th, 10th, 11th, 12th, 13th, 15th, 17th, 18th and 20th. The result is obtained from similarity of different venue information that we obtained from Foursquare API.

4-2. Price of apartment

If you want to move to Paris and you are looking for an apartment in a residential area. However, it seems to be too expensive to take an apartment in the 6th or the 7th district. (-14 590 euros per m2 in the 6th district) Is there any neighborhood which is similar to these districts ? With clustering, we found out that there are other districts like 14th, 16th and 19th which are similar to 6th and 7th districts. And with the dataset merged, we know that we can find an apartment less expensive in the 19th district than in the 6th district. In conclusion, there are other districts which have similar characteristics, but which have advantage in the price.

	District	Price/m2	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Co
0	75001	13 940	48.859	2.342	0	Café	Hotel	French Restaurant	Clothing Store	Plaza	Tea Room	Wine Bar	Historic Site	Rest
1	75002	12 250	48.865	2.343	0	French Restaurant	Wine Bar	Bakery	Cocktail Bar	Japanese Restaurant	Italian Restaurant	Cheese Shop	Sandwich Place	Optica
2	75003	12 650	48.864	2.361	0	Wine Bar	Japanese Restaurant	Art Gallery	French Restaurant	Coffee Shop	Seafood Restaurant	Italian Restaurant	Bakery	Sar
3	75004	13 660	48.860	2.351	0	Bakery	Hotel	Art Gallery	French Restaurant	Plaza	Park	Historic Site	Coffee Shop	Ice
4	75005	12 750	48.845	2.347	0	French Restaurant	Hotel	Café	Pub	Italian Restaurant	Bar	Indie Movie Theater	Bakery	Wi
5	75006	14 590	48.849	2.330	1	French Restaurant	Hotel	Bakery	Fountain	Wine Bar	Chocolate Shop	Garden	Ice Cream Shop	Rest
6	75007	14 100	48.856	2.321	1	French Restaurant	Hotel	Café	Bakery	Coffee Shop	Bistro	Italian Restaurant	Art Museum	Jap Rest
7	75008	11 530	48.876	2.318	0	Hotel	French Restaurant	Italian Restaurant	Café	Thai Restaurant	Bar	Bakery	Wine Shop	T
8	75009	11 070	48.872	2.340	0	Hotel	French Restaurant	Pedestrian Plaza	Coffee Shop	Creperie	Cocktail Bar	Italian Restaurant	Indie Movie Theater	Gc
9	75010	10 480	48.871	2.356	0	French Restaurant	Cocktail Bar	Breakfast Spot	Hotel	Seafood Restaurant	Salad Place	Art Gallery	Bakery	Rest
10	75011	10 820	48.857	2.380	0	French Restaurant	Bar	Bistro	Cocktail Bar	Japanese Restaurant	Café	Pastry Shop	Italian Restaurant	E
11	75012	10 110	48.841	2.388	0	Hotel	Bistro	French Restaurant	Supermarket	Chinese Restaurant	Café	Garden	Sporting Goods Shop	Furn Home
12	75013	9 640	48.832	2.356	0	Thai Restaurant	Hotel	Bakery	Vietnamese Restaurant	French Restaurant	Indian Restaurant	Italian Restaurant	Bar	Rest
13	75014	10 680	48.833	2.326	1	French Restaurant	Hotel	Bar	Italian Restaurant	Bakery	Bistro	Restaurant	Outdoor Sculpture	G
14	75015	10 700	48.841	2.300	0	French Restaurant	Italian Restaurant	Hotel	Bakery	Coffee Shop	Korean Restaurant	Lebanese Restaurant	Bar	Superr
15	75016	11 480	48.864	2.277	1	French Restaurant	Bakery	Café	Italian Restaurant	Hotel	Hotel Bar	Pizza Place	Chinese Restaurant	Rest
16	75017	11 020	48.883	2.322	0	French Restaurant	Wine Bar	Hotel	Bar	Restaurant	Italian Restaurant	Thai Restaurant	Pizza Place	Vietn Rest
17	75018	10 180	48.892	2.344	0	French Restaurant	Café	Bar	Italian Restaurant	Theater	Wine Bar	Pizza Place	Seafood Restaurant	Rest
18	75019	9 040	48.882	2.382	1	French Restaurant	Pool	Bus Stop	Italian Restaurant	Bar	Bakery	Supermarket	Restaurant	Bike R Bike
19	75020	9 450	48.865	2.398	0	French Restaurant	Bar	Bakery	Bistro	Japanese Restaurant	Wine Bar	Park	Supermarket	

5. Discussion

This time we set the value of cluster as 2 for dividing the districts by 2 : residential area and commercial area. However, if we make more clusters, we may get the results more specified and clarified. It will help us to define and to understand the characteristics of different districts in Paris.

6. Conclusion

People are frequently moving into new cities. And when people move into a new city, the price of an apartment could be important. Having a neighborhood recommendation based on location data and on price can serve to be an impressive tool to better organise a city ressources.